CLAIMS

What is claimed is:

1. A method comprising:

receiving, into a capacity planning system, workload information representing an expected workload of client accesses of streaming media files from a site;

receiving, into said capacity planning system, at least one service parameter that defines a desired service characteristic to be provided by a media server configuration under the expected workload; and

determining, by said capacity planning system, for at least one server configuration, how many servers of said at least one server configuration to be included at said site for supporting the expected workload in compliance with said at least one service parameter.

2. The method of claim 1 wherein said at least one service parameter comprises at least one performability parameter that defines a desired service characteristic to be provided by a media server configuration during periods of degraded service under the expected workload.

3. The method of claim 2 wherein said at least one performability parameter specifies a limit on the amount of degradation of service encountered during said periods of degraded service.

4. The method of claim 2 wherein said at least one performability parameter comprises at least one selected from the group consisting of:

a regular-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods of performance degradation under regular system operation of a media server configuration, and

a node-failure-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods in which one or more nodes of a clustered media server configuration have failed.

5. The method of claim 2 wherein said at least one performability parameter comprises a regular-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods of performance degradation under regular system operation of a media server configuration, and a node-failure-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods in which one or more nodes of a clustered media server configuration have failed.

6. The method of claim 1 wherein said at least one service parameter comprises at least one basic capacity parameter.

7. The method of claim 6 wherein said at least one basic capacity parameter comprises at least one selected from the group consisting of:

a statistical demand guarantee that specifies a desired limit on the percentage of time that a media server configuration is overloaded under the expected workload, and

a utilization constraint that specifies a desired limit on the percentage of time that a media server configuration is at or near its capacity under the expected workload.

8. The method of claim 6 wherein said at least one basic capacity parameter comprises a statistical demand guarantee that specifies a desired limit on the percentage of time that a media server configuration is overloaded under the expected workload, and a utilization constraint that specifies a desired limit on the percentage of time that a media server configuration is at or near its capacity under the expected workload.

9. The method of claim 6 wherein said at least one service parameter further comprises at least one performability parameter that defines a desired limit on the amount of degradation of service encountered during said percentage of time that a media server configuration is overloaded under the expected workload.

10. The method of claim 6 wherein said at least one service parameter further comprises at least one performability parameter that defines a desired limit on the amount of continuous overload encountered at any given time by a media server configuration under the expected workload.

11. A method comprising:

receiving, into a capacity planning tool, information about a first server configuration;

receiving, into said capacity planning tool, workload information representing an expected workload of client accesses of streaming media files from a site;

receiving, into said capacity planning system, at least one performability parameter that defines a desired service characteristic to be provided by a media server configuration during non-compliant periods of operation under the expected workload; and

said capacity planning tool determining how many servers of said first server configuration to be included at said site for supporting the expected workload in compliance with said at least one performability parameter.

12. The method of claim 11 wherein said non-compliant periods of operation comprise periods of degraded performance in servicing said expected workload.

13. The method of claim 12 wherein said degraded performance is performance in which said media server configuration is unable to satisfy real-time constraints of at least one stream being served.

14. The method of claim 12 wherein said degraded performance is performance in which said media server configuration is unable to serve at least one stream so as to avoid interruptions in the presentation of such stream.

15. The method of claim 12 wherein said degraded performance results from overload of said media server configuration.

16. The method of claim 11 wherein said non-compliant periods of operation comprise periods of at least one node failure of a clustered media server configuration.

17. The method of claim 11 further comprising:

receiving, into said capacity planning system, at least one basic capacity parameter that defines a desired service characteristic to be provided by a media server configuration during compliant periods of operation under the expected workload.

18. The method of claim 17 wherein said compliant periods of operation comprise periods in which said media server configuration is not overloaded under the expected workload.

19. The method of claim 17 further comprising:

said capacity planning tool performing basic capacity planning to determine how many servers of said first server configuration to be included at said site for supporting the expected workload in compliance with said at least one basic capacity parameter.

20. The method of claim 19 further comprising:

said capacity planning tool determining how many servers of said first server configuration to be included at said site for supporting the expected workload in compliance with said at least one basic capacity parameter and said at least one performability parameter.

21. The method of claim 11 wherein said at least one performability parameter comprises at least one selected from the group consisting of:

a regular-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods of performance degradation under regular system operation of said media server configuration, and

a node-failure-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods in which one or more nodes of a clustered media server configuration have failed.

22. A method comprising:

receiving, into a capacity planning tool, workload information representing an expected workload of client accesses of streaming media files over a period of time T;

said capacity planning tool determining, for at least one media server configuration under evaluation, an amount of overload encountered by said at least one media server configuration during each of a plurality of time intervals of said expected workload.

23. The method of claim 22 where each of said plurality of time intervals have a size I where $I < T$.

24. The method of claim 22 wherein beginning points of each of said plurality of time intervals are separated by a Step amount.

25. The method of claim 24 wherein said Step $< I$.

26. The method of claim 24 wherein each of said intervals has a duration of 1 hour and said Step is 1 minute.

27. The method of claim 22 wherein said capacity planning tool receives at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by a media server configuration under the expected workload.

28. The method of claim 27 wherein said capacity planning tool evaluates said amount of overload encountered by said at least one media server configuration during each of said plurality of time intervals to determine whether said at least one media server configuration satisfies said at least one performability parameter.

29. The method of claim 27 wherein said at least one performability parameter comprises at least one selected from the group consisting of:

a regular-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods of performance degradation under regular system operation of a media server configuration, and

a node-failure-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods in which one or more nodes of a clustered media server configuration have failed.

30. A method comprising:

receiving, into a capacity planning tool, workload information identifying an expected workload of client accesses of streaming media files from a server over a period of time T;

determining, by said capacity planning tool, an interval overload profile for a media server configuration under evaluation, wherein said interval overload profile specifies an amount of overload of said media server configuration for each of a plurality of time intervals of duration I of said expected workload, where $I < T$; and

said capacity planning tool determining based at least in part on the interval overload profile whether said media server configuration under evaluation supports the expected workload in compliance with defined service parameters that define service characteristics desired by a service provider.

31. The method of claim 30 wherein beginning points of each of said plurality of time intervals are separated by a Step amount.

32. The method of claim 31 wherein said Step $< I$.

33. The method of claim 30 wherein said defined service parameters include at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by a media server configuration under the expected workload.

34. The method of claim 33 wherein said capacity planning tool evaluates said interval overload profile for said media server configuration under evaluation to determine whether said media server configuration under evaluation satisfies said at least one performability parameter.

35. The method of claim 33 wherein said at least one performability parameter comprises at least one selected from the group consisting of:

a regular-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods of performance degradation under regular system operation of a media server configuration, and

a node-failure-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods in which one or more nodes of a clustered media server configuration have failed.

36. A system comprising:

means for receiving workload information representing an expected workload of client accesses of streaming media files from a site over a period of time T; and

means for determining, for at least one media server configuration under evaluation, an amount of overload encountered by said at least one media server configuration during servicing each of a plurality of time intervals of said expected workload.

37. The system of claim 36 further comprising:

means for receiving information specifying duration of each of said time intervals.

38. The system of claim 36 where each of said plurality of time intervals have a duration I where $I < T$.

39. The system of claim 36 wherein beginning points of each of said plurality of time intervals are separated by a Step amount.

40. The system of claim 39 wherein said Step is smaller in duration than a duration I of each of said intervals.

41. The system of claim 36 further comprising:

means for receiving at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by said at least one media server configuration under evaluation under the expected workload.

42. The system of claim 41 further comprising:

means for evaluating the determined amount of overload encountered by said at least one media server configuration under evaluation for each of said plurality of time intervals to determine whether said at least one media server configuration under evaluation satisfies said at least one performability parameter.

43. The system of claim 41 wherein said at least one performability parameter comprises at least one selected from the group consisting of:

a regular-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods of performance degradation under regular system operation of said at least one media server configuration under evaluation, and

a node-failure-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods in which one or more nodes of a clustered media server configuration under evaluation have failed.

44. A system comprising:

a media profiler operable to receive workload information for a service provider's site and generate a workload profile for a server configuration under consideration for supporting the service provider's site; and

a capacity planner operable to receive the generated workload profile for the server configuration under consideration and determine how many servers of said server configuration are needed to provide a media server solution having sufficient capacity for supporting the site's workload in compliance with defined performability parameters that specify a desired limit on degradation of quality of service provided by said media server solution during periods of degraded service.

45. The system of claim 44 wherein said periods of degraded service is periods in which said media server configuration is unable to serve at least one stream so as to avoid interruptions in the presentation of such stream.

46. The system of claim 44 wherein said defined performability parameters comprise at least one selected from the group consisting of:

a regular-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods of degraded service under regular system operation of said media server solution, and

a node-failure-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods in which one or more nodes of a clustered media server solution have failed.